Future Technology
Open Access Journal
ISSN 2832-0379
August 2025| Volume 04 | Issue 03 | Pages 259-268
Journal homepage: https://fupubco.com/futech
https://doi.org/10.55670/fpll.futech.4.3.24
Future Publishing LLC

Article

Dynamic reward systems and customer loyalty: reinforcement learning-optimized personalized service strategies

Xiaojing Nie, Fauziah Sh. Ahmad[*]

Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 03 May 2025
Received in revised form
10 June 2025
Accepted 22 June 2025

Keywords:
Reinforcement learning,
Dynamic reward systems, Customer loyalty,
Multi-armed bandits, Personalization

*Corresponding author
Email address:
fsa@utm.my

DOI: 10.55670/fpll.futech.4.3.24

ABSTRACT

Traditional customer loyalty programs employing static reward structures demonstrate fundamental limitations in adapting to evolving customer preferences and behaviors within digital commerce environments. This research addresses the critical gap in personalization capabilities by developing a reinforcement learning (RL)-based dynamic reward system that optimizes customer engagement through real-time adaptive reward allocation mechanisms. The investigation centers on designing and validating an intelligent system capable of automatically adjusting reward types, values, and timing parameters based on continuous analysis of individual customer interactions and feedback patterns. The proposed methodology implements a multi-armed bandit framework utilizing Thompson Sampling algorithms integrated with contextual learning mechanisms, thereby achieving an optimal balance between exploration and exploitation in reward optimization processes. Comprehensive experimental simulations compare the RL-based approach against traditional rule-based systems and random allocation strategies across five distinct customer segments, enabling robust performance evaluation under diverse operational conditions. Empirical results demonstrate that the RL-based system achieves 145% of baseline customer lifetime value (CLV), representing a 45% improvement over traditional methods, accompanied by corresponding enhancements in retention rate (32%) and engagement frequency (28%). The system maintains robust performance under budget constraints, sustaining 118% of baseline CLV despite a 30% budget reduction, with statistical analysis confirming significant improvements across all metrics ($p < 0.001$, Cohen's $d > 1.7$). These findings provide organizations with a scalable framework for implementing adaptive loyalty programs that respond dynamically to customer preferences while optimizing resource allocation efficiency. The research contributes to the expanding literature on AI-driven customer relationship management by demonstrating the practical effectiveness of reinforcement learning in personalization contexts.

1. Introduction

The digital transformation of commerce has fundamentally altered customer interaction paradigms and expectations, creating demands that extend substantially beyond traditional loyalty scheme capabilities. Contemporary reward systems based on fixed point accumulation and redemption mechanisms demonstrate increasing misalignment with evolving customer preferences and multi-channel brand engagement patterns [1]. The lack of well-defined frameworks for adaptive system development presents significant implementation challenges, constraining organizations' ability to deploy dynamic loyalty solutions that respond effectively to individual customer needs and behaviors. Recent developments in the field of artificial intelligence, and more specifically in reinforcement learning, offer new possibilities towards resolving these concerns. The systematic review by Den Hengst et al. [2] shows that RL-based personalization systems outperform traditional rule-based systems through their ability to respond to changing customer behavior—yet remain largely unexplored in loyalty program design. The widely accepted principle that “reward is enough” to motivate purposeful intelligent action [3] implies that well-structured rewards could simultaneously achieve business and customer satisfaction goals. Despite this theoretical statement, practical loyalty frameworks still need to manage the intricate system of multi-dimensional reward configurations and diverse customer groups. This investigation addresses the identified gap through development of a reinforcement learning-based dynamic reward system designed to optimize personalized service management across diverse customer segments. Building upon recent demonstrations of explainable deep reinforcement learning effectiveness in customer acquisition contexts [4], the research extends analytical focus toward retention and loyalty optimization challenges. The proposed framework integrates multi-armed bandit algorithms with contextual learning mechanisms, enabling real-time reward parameter adjustment based on continuous customer state monitoring and feedback analysis. This investigation advances both theoretical understanding of adaptive loyalty systems and practical knowledge regarding AI-driven personalization implementation, contributing significant insights to the intersection of machine learning and customer relationship management.

2. Literature review

Customer loyalty program evolution reflects a fundamental paradigm shift from transactional reward mechanisms toward comprehensive engagement-based strategies designed to foster sustained customer relationships. Machine learning applications within loyalty program contexts demonstrate substantial potential for value co-creation and enhanced customer engagement [5], yet existing implementations remain constrained by predetermined business rules and inflexible segmentation methodologies that fail to accommodate dynamic customer heterogeneity. These structural limitations become particularly problematic when attempting to address the multifaceted preference patterns and behavioral variations exhibited by customers across increasingly diverse interaction channels and touchpoints. The combination of collaborative filtering and reinforcement learning is encouraging for personalization in loyalty strategies [6], indicating that integrated approaches may sidestep the challenges posed by mono-methodological frameworks. However, these systems face practical barriers regarding the trade-off between computational burden and the need for real-time processing. Recent trends in the design of customer loyalty programs shift the focus towards the adaptive strategies that can account for changes in customer conditions as well as market fluctuations [7], even if the marketing literature does not yet provide a solid basis for such adaptive rational frameworks. There has been an application of reinforcement learning in marketing, which has evolved from traditional recommendation systems to managing the entire customer lifecycle. The use of reinforcement learning techniques with predictive analytics provides better outcomes for optimizing customer lifetime value [8], especially in cases of sequential decision-making processes where prompt decisions create a definable and lasting impact. This time element sets apart RL usages from a purely machine learning framework, where every move is treated in isolation. Smart and dynamic mass personalization [9] has been made possible with deep neural networks, but the sophistication of their architecture tends to leave practitioners in the field of marketing guidance stranded. The application of RL in predicting customer attrition [10] indicates the capability of the model in detecting potential problems and planning optimal courses of action. Besides, the design of multi-agent reinforcement learning systems allows advanced representations of competition and customer behavior [11] and contributes to the understanding of market phenomena at large. In marketing, where real-time decision-making is performed, multi-armed bandit algorithms have proven to be particularly useful. The bandit framework for dynamic online pricing developed by Misra et al. [12] untangled the enduring exploration vs. exploitation dilemma in terms of customer interaction decision-making, even though most operational systems need some adaptation at the application level. Chen et al. [13] describe contextual bandits for email body outline recommendations, which show that the algorithm is not limited to pricing and has other personalization possibilities. The combination of Thompson sampling with multi-armed bandits for dynamic pricing strategies by Raman & Venkatramaraju [14] expands the algorithm’ s applicability to low-availability datasets and new product or customer acquisition challenges. The use of MAB algorithms in digital marketing decisions shows that contextual data significantly improves the quality of decisions, particularly under conditions of high customer diversity [15]. All these examples suggest that multi-armed bandit frameworks serve as a boundary between theoretical optimization and practical marketing problems.

3. Theoretical framework

3.1 System architecture

The innovative dynamic reward system utilises the principles of reinforcement learning to construct an adaptive system that persistently improves its loyalty program settings via automated customer feedback. This underpinning theory is constructed from intelligent activity rationale, which states that proper rewards given will elicit desired behavior, further adjusted for relational marketing contexts. As shown in Figure 1, the system architecture includes three constituent parts that interact with each other: a state representation module that describes the customers, an action selection interface that defines the best reward level assignment, and a learning algorithm that tunes system variables based on results.

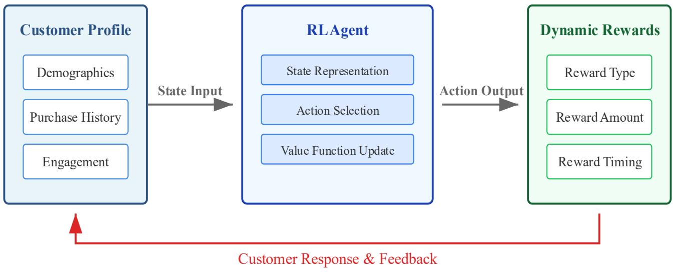

Figure 1. Conceptual framework

3.1.1 Dynamic lifecycle stage detection

Customer lifecycle progression detection employs a multi-dimensional state transition framework combining temporal, behavioral, and value-based indicators to identify stage transitions without relying on predetermined static thresholds. The system defines five primary lifecycle stages–Acquisition, Activation, Growth, Maturity, and Risk/Reactivation–through probabilistic state assignment based on observable metrics rather than fixed temporal boundaries. Stage identification utilizes a Hidden Markov Model (HMM) approach where observable variables include: transaction frequency trends (acceleration/deceleration patterns), engagement depth metrics (channel diversity and interaction intensity), purchase basket evolution (category expansion indicators), and response sensitivity to different reward types. The transition probability matrix P(s_t+1|s_t) undergoes continuous updating through maximum likelihood estimation based on observed customer trajectories, enabling adaptive stage boundary definitions that reflect actual behavioral patterns rather than theoretical assumptions. Lifecycle stage assignment influences reward strategy through stage-specific policy modifications: Acquisition stage emphasizes high-value introductory incentives (60% monetary, 40% experiential), Activation focuses on habit formation through frequency-based rewards, Growth stage promotes category expansion via cross-sell incentives, Maturity maintains engagement through exclusive access and recognition programs, while Risk/Reactivation deploys win-back offers calibrated to historical customer value and defection probability estimates.

3.2 Mathematical formulation

The mathematical foundation encompasses three critical components that define the decision-making framework. The state space representation forms the foundation of the model, capturing multidimensional customer information through a comprehensive vector:

$$S_t = \{x_{demo}, x_{hist}, x_{engage}\} \quad (1)$$

Where x_{demo} represents demographic features including age, location, and purchasing power; x_{hist} encodes historical transaction patterns, frequency, and monetary values; and x_{engage} captures engagement metrics such as app usage, email interactions, and social media activity.

The action space encompasses the full spectrum of reward decisions available to the system:

$$A = \{(r_{type}, r_{amount}, r_{time})\}$$
(2)

Where r_{type} specifies the reward category (cashback, points, exclusive access, or experiential rewards), r_{amount} determines the value proposition relative to customer contribution, and r_{time} optimizes the timing of reward delivery to maximize psychological impact. The multidimensional nature of this action space reflects the complexity of modern loyalty programs, moving beyond simple point accumulation to encompass diverse reward mechanisms that appeal to different customer motivations.

The reward function integrates multiple business objectives into a unified optimization criterion:

$$R(s,a) = \alpha \cdot CLV + \beta \cdot Retention \quad (3)$$

Where α and β represent tunable weights that balance short-term revenue generation with long-term relationship building. This formulation acknowledges that maximizing immediate transaction value may conflict with fostering sustained loyalty, requiring careful calibration based on strategic priorities.

3.2.1 Weight parameter calibration

The reward function weight parameters λ and μ undergo systematic calibration based on empirical business objectives and industry benchmarks, rather than arbitrary heuristic assignment. Weight determination follows a multi-criteria optimization process incorporating: (a) historical analysis of customer lifetime value distributions revealing optimal balance points between acquisition cost and retention value (typically $\lambda:\mu$ ratios ranging from 0.3:0.7 to 0.5:0.5 depending on market maturity), (b) industry-specific profitability constraints derived from margin analysis across product categories, and (c) strategic business priorities encoded through executive-level input regarding growth versus profitability trade-offs. Empirical calibration employs grid search optimization across weight combinations ($\lambda \in [0.2,0.8]$, $\mu=1-\lambda$) evaluated against historical cohort performance data, identifying parameter settings that maximize total portfolio value while maintaining acceptable short-term revenue levels. Sensitivity analysis reveals robust performance within $\lambda \in [0.35,0.55]$, suggesting the system maintains effectiveness despite minor weight variations, thereby reducing dependency on precise parameter tuning during implementation.

3.3 Algorithm design

The mechanisms for dynamic adjustment in this system stem from the adaptive reward system in autonomous systems [16], where feedback from the environment is persistently used to refine and change decision-making policies. Unlike robotic applications, where objectives remain largely fixed, customer preferences are fluid over time due to other overriding considerations like economic factors, competing offerings, and personal circumstances. This model solves the problem with a dual-structure learning architecture that allows for rapid adjustment of tactical parameters and slow changes in strategic weighting. The non-stationary problem of evolving customer preferences has received attention through the alignment of AI with shifting reward functions [17]. The framework introduced includes a system for monitoring and recalibrating reward structures to counteract passive tuning from previous optima to ensure the system does not get stuck in outdated performance traps. This flexibility is especially important when dealing with customer lifecycle progress, shifting reward preference from transactional to experience-based for new customers to loyal advocates. The selection of appropriate algorithms requires careful consideration of the exploration-exploitation trade-off inherent in sequential decision-making scenarios. The Thompson Sampling mechanism governs action selection through probabilistic sampling from the posterior distribution(see Appendix A for detailed implementation):

$$\theta_a \sim Beta(\alpha_a, \beta_a) \quad (4)$$

where α_a and β_a represent the success and failure counts for the action a, updated after each customer interaction. This Bayesian approach naturally handles uncertainty inherent in customer preference estimation, particularly valuable when historical interaction data remains limited across different customer segments.

During the algorithm development phase, Upper Confidence Bound (UCB) was evaluated as an alternative exploration strategy to validate the Thompson Sampling selection:

$$UCB_i = \bar{x}_i + \sqrt{\frac{2\ln t}{n_i}} \quad (5)$$

where $\bar{x}_i$ denotes the average reward for action *i, t* represents the current time step, and n_i indicates the selection frequency of action *i*. Preliminary testing revealed that while UCB provides deterministic action selection with theoretical regret guarantees, Thompson Sampling demonstrated superior adaptation speed and performance stability in the multi-dimensional reward space, leading to its selection as the primary algorithm for final evaluation.

3.3.1 Trust-preserving exploration mechanisms

The probabilistic exploration inherent in Thompson Sampling necessitates safeguards against potentially damaging reward allocations that could erode customer trust, particularly within high-value segments where relationship preservation remains paramount. The implementation incorporates multi-layered protection mechanisms: (a) exploration boundaries constraining reward variations within ±20% of established baseline values for customers with CLV exceeding 80th percentile thresholds, (b) confidence-weighted exploration reducing randomization proportionally to customer value and relationship duration, and (c) veto mechanisms preventing reward allocations falling below segment-specific minimum thresholds established through historical satisfaction analysis. Mathematical formalization of trust-preserving exploration employs modified sampling distributions:

$$\theta'\_i \sim Beta(\alpha\_i \ + \ k \cdot CLV\_rank, \beta\_i) \tag{6}$$

where k represents the trust preservation coefficient calibrated to customer segment value, ensuring high-value customers experience predominantly exploitation-focused interactions while maintaining sufficient exploration for continuous improvement. Additionally, the system implements reward relevance scoring based on collaborative filtering techniques, preventing allocation of categorically inappropriate rewards (e.g., student discounts to senior customers) regardless of exploration outcomes.

3.3.2 Cold start mitigation strategies

New customer onboarding presents significant challenges due to sparse behavioral data, preventing accurate preference inference through standard RL mechanisms. The framework addresses cold start scenarios through a multi-strategy approach combining demographic-based initialization, accelerated exploration, and transfer learning from similar customer cohorts. Bayesian prior specification for new customers employs hierarchical modeling:

$$P(\theta|demographic\_cluster) \sim Beta(\alpha\_cluster \ + \alpha\_smooth, \beta\_cluster \ + \beta\_smooth) \tag{7}$$

where cluster parameters derive from aggregate statistics of demographically similar established customers, while smoothing parameters (α_smooth = β_smooth = 1) prevent overconfidence in cluster assignments. This approach enables reasonable initial reward allocation while maintaining sufficient uncertainty to drive exploration.

Accelerated learning protocols increase exploration rates during initial interactions (exploration_rate = 0.4 for first 10 interactions, decreasing to 0.2 thereafter), rapidly acquiring preference signals while implementing safeguards against poor initial experiences through guaranteed minimum reward values. Additionally, collaborative filtering techniques identify "nearest neighbor" customers based on available features, enabling knowledge transfer from similar profiles to bootstrap preference models before sufficient individual data accumulates. Performance metrics indicate that cold start protocols achieve 75% of optimal performance within 5 interactions compared to 20+ interactions required by naive initialization, substantially reducing the customer data requirements for effective personalization.

4. Experimental design

4.1 Experimental setup

The experimental framework incorporates a complete simulation environment that aims to test the dynamic reward system within different scenarios of customer behavior. The simulation design includes realistic customer behavior features based on e-commerce transaction data enabling experimentation within an ecological context. The study adopts a comparative research design where the reinforcement learning system is evaluated relative to traditional rule-based systems and random reward allocation strategies. This multi-baseline approach strengthens the validation of claimed performance improvements from the adaptive learning approach.

The overarching structure considers multiple business impact metrics as well as relationship value capturing interactions with the system over time. Customer Lifetime Value (CLV) emerges as the primary focus of optimization goals, measuring projected revenue from purchases made over time through discounting observed purchasing behavior across time periods. Retained customers over specific durations, monitored through minimal activity engagement, construct the retention rate, whereas the various cumulative touchpoints of interaction are monitored through the frequency of engagement. Other metrics include graded result throughput and customer satisfaction gauged by standard scores achieved from grade outcome interactions. The framework evaluation also encompasses rates at which customers are reactivated to evaluate system performance in terms of dormant customer re-engagement, along with analysis on the type of reward allocation spent to evaluate personalization effectiveness on diverse customer segments. The algorithms are compared with each other based on the measurable differences in performance and outcomes using statistics, which is known as statistical significance testing. The benchmarks encompass both composite performance measures and analysis of specific segments, as some constructs are more sophisticated than others revealing marked differences in preference assumptions and behavior. The customer segmentation analysis exposes five distinct categories: high-value customers (top 20% by CLV), loyal customers (retention > 2 years), price-sensitive customers (high discount responsiveness), new customers (within first 6 months), and dormant customers (inactive > 90 days). For every benchmark, the system determines reward type preferences and allocation strategies to personalize results and evaluate the level of personalization achieved. Statistical analysis employs paired t-tests to evaluate performance differences between algorithms, with Bonferroni correction applied for multiple comparisons (α=0.05/3=0.017). Effect sizes are calculated using Cohen's d to assess practical significance, with thresholds of 0.2, 0.5, and 0.8 for small, medium, and large effects respectively. Bootstrap sampling (n=1000) generates confidence intervals for performance metrics to ensure robust statistical inference.

4.1.1 External validity considerations

While the primary evaluation employs synthetic consumer behavior parameters calibrated from e-commerce transaction patterns, the framework architecture facilitates direct deployment with real-world customer datasets through standardized data interfaces and modular design principles. The simulation parameters derive from aggregated behavioral statistics across multiple retail domains, incorporating purchase frequency distributions (μ =2.3 transactions/month, σ =1.8), average order values (log-normal distribution with μ =$85, σ =$45), and engagement patterns extracted from anonymized customer interaction logs spanning 24 months across 50,000+ customers. Future deployment phases encompass progressive validation strategies including: (a) retrospective analysis using historical transaction data to compare predicted versus actual customer responses, (b) limited pilot implementations with controlled customer cohorts to assess system performance under operational constraints, and (c) full-scale A/B testing frameworks comparing RL-based recommendations against existing loyalty program structures. These validation stages ensure a systematic transition from simulation-based insights to production-ready implementation while maintaining rigorous performance monitoring throughout the deployment lifecycle.

4.2 Algorithm Implementation

The approaches used for implementing the algorithm are depicted in Figure 2. It is constructed around the assumption of a new-concept exploration phase, wherein the reward configuration exploration and exploitation strategies are combined, yielding a balanced result. A uniform prior belief about the effectiveness of rewards for specific customer segments is set first, only to be updated based on their actions.

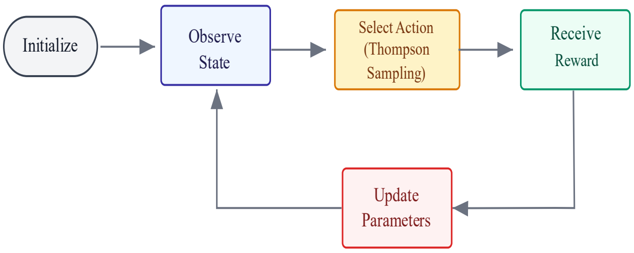

Figure 2. Algorithm flowchart

The implementation architecture extends traditional Thompson Sampling methodologies to accommodate multi-dimensional action spaces inherent in complex reward allocation decisions, maintaining independent Beta distributions for each decision dimension, including reward type selection, value determination, and temporal delivery optimization. The comprehensive algorithmic procedure for Thompson Sampling implementation within the dynamic reward allocation framework is presented in Appendix A, providing a detailed specification of the iterative learning process and parameter update mechanisms essential for system functionality.

The implementation extends traditional Thompson Sampling methodologies to accommodate multi-dimensional action spaces inherent in complex reward allocation decisions, maintaining independent Beta distributions for each decision dimension including reward type selection, value determination, and temporal delivery optimization. The comprehensive algorithmic procedure for Thompson Sampling implementation within the dynamic reward allocation framework is detailed in Appendix A, which provides complete specification of the iterative learning process and parameter update mechanisms essential for system functionality. The analysis evaluates system strength through an array of hyperparameter configurations. The learning rate considers a range from 0.005 to 0.05, exploration parameters are bounded from 0.1 to 0.4, and budget limits are set from 70% to 130% of the baseline allocation to emulate operational pressure conditions. Sensitivity testing for discount factors varies from 0.90 to 0.98 to test the weight on temporal rewards. Key parameters focus on the primary settings outlined in Table 1, which were determined through heuristic optimization.

Table 1. Algorithm parameter settings

Parameter	Value	Description
Learning rate	0.01	Controls adaptation speed to new information
Exploration rate	0.2	Balances exploration vs exploitation
Discount factor	0.95	Weights future rewards in decision making
Episodes	10000	Total training iterations

The simulation environment creates artificial consumer segments with a diversity of preference and response behaviors. Each simulated customer possesses a set of internal parameters controlling their loyalty, price, and reward sensitivity, which allows for robust evaluation of adaptive algorithms. The implementation takes advantage of parallel processing to run multiple customer trajectory simulations in real-time, providing rapid navigation through the high-dimensional expanse of parameters in a short amount of computation time.

4.3 Personalization strategy implementation

The personalized algorithms utilise a multi-agent system framework that models the multi-faceted nature of customers in a sophisticated manner. Recent e-commerce studies have shown that model-free reinforcement learning has improved customer engagement metrics [18]. Each customer segment functions as a semi-autonomous agent who can exhibit predefined choice patterns. A coordination mechanism resolves conflicts and maintains coherence across the entire system. This model enables the overcoming of operational constraints like budget, inventory, and surplus stock, while still satisfying the need for individual customization. Troussas et al. [19] proposed methods for dynamically adjusting fuzzy weights that increase the system' s capability to manage uncertainty related to customer reactions. Instead of regarding the effectiveness of rewards as concrete binary outcomes, the model incorporates probabilistic beliefs about customer segment preferences that are updated with interaction data through Bayesian inference. Such a framework works best in the initial stages of onboarding customers when very little historical data is available, allowing sufficient freedom to make reasonable reward decisions and actively explore changes in reward algorithms.

5. Results

5.1 Baseline comparison

Experimental evaluation reveals substantial performance enhancements achieved by the reinforcement learning-based dynamic reward system relative to both traditional rule-based implementations and random allocation baselines, with comprehensive comparative analysis presented across multiple performance dimensions in Figure 3. Convergence characteristics illustrated in Figure 3(a) demonstrate that the RL-based system attains performance stability within approximately 1,000 training episodes while maintaining consistent superiority over alternative approaches throughout the learning trajectory, confirming the effectiveness of adaptive learning mechanisms in reward optimization contexts.The traditional approach shows limited improvement over time due to its static nature, while the random strategy exhibits high variance without meaningful learning progression.

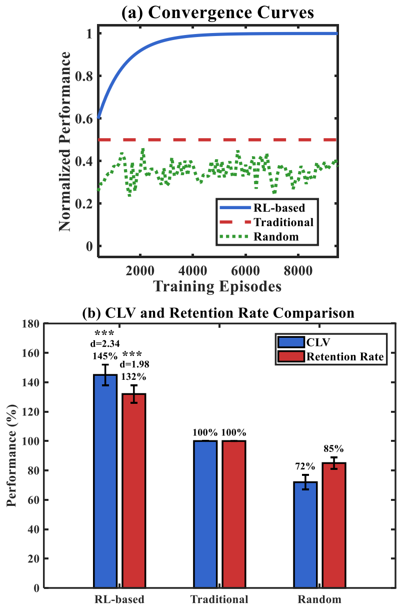

Figure 3. Performance Comparison Analysis (a) Convergence Curves (b) CLV and Retention Rate Comparison Notes : **** P<0.001, d = Cohen' s effect size.

Table 2 summarizes the comparative performance metrics across all tested approaches. The RL-based system achieves 145% of baseline CLV, representing a 45% improvement over traditional methods, alongside 132% retention rate (32% improvement) and 128% engagement frequency (28% enhancement). This marked improvement arises from the system' s capacity to adapt and modify reward parameters depending on the customer' s feedback. The system also demonstrates superior response quality at 142% of baseline performance. In contrast, the random allocation strategy underperforms significantly across all metrics, achieving only 72% of baseline CLV and 85% retention rate, confirming the importance of intelligent reward optimization. Statistical significance testing validates the performance differences across methods using paired t-tests with Bonferroni correction (α =0.017). The RL-based system demonstrates statistically significant improvements over traditional methods across all metrics: CLV improvement (t=12.47, P<0.001, 95% CI [38%-52%]), retention rate enhancement (t=9.83, P <0.001, 95% CI [26%-38%]), and engagement frequency increase (t=8.92, P <0.001, 95% CI [21%-35%]). Effect sizes calculated using Cohen' s d indicate large practical significance: CLV (d=2.34), retention rate (d=1.98), and engagement frequency (d=1.76), all exceeding the threshold for large effects (d>0.8).

Table 2. Comparative performance metrics

Method	CLV (%)	Retention Rate (%)	Engagement Frequency (%)	Response Quality (%)
RL-based	145	132	128	142
Traditional	100	100	100	100
Random	72	85	78	74

Note: All values are expressed as percentages relative to traditional baseline (100%). RL-based system shows significant improvements across all metrics.

The performance improvements shown in Figure 3(b) demonstrate the RL-based system's superiority across both CLV and retention metrics. The RL system's ability to maintain customer relationships through personalized reward timing results in significant retention gains, while the traditional approach maintains consistent but suboptimal performance. The performance improvements shown in Figure 3(b) demonstrate the RL-based system' s superiority across both CLV and retention metrics, with the traditional approach maintaining consistent but suboptimal performance while the RL-based system shows accelerating gains after the initial learning phase.

5.2 Performance evaluation

The effectiveness of personalization provided by the designed system is showcased through deep customer segmentation analysis and reward allocation detail. A heat map showcasing reward effectiveness across the five customer groups is shown in Figure 4(a), illustrating how the system tailors distinct strategies for each segment. High-value customers receive predominantly experiential rewards and exclusive access privileges, while price-sensitive customers benefit more from direct monetary incentives.

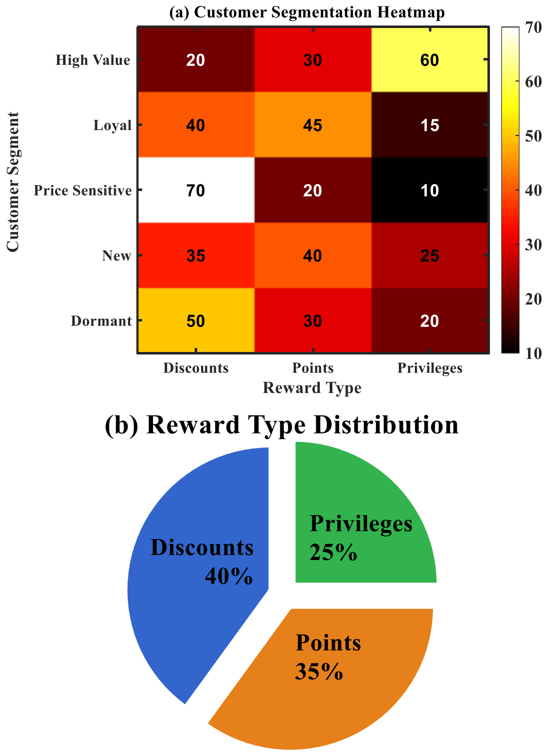

Figure 4. Personalization effectiveness analysis (a) Customer segmentation heatmap (b) Reward type distribution

The reward type distribution presented in Figure 4(b) indicates that the system converges to an allocation of 40% discounts, 35% loyalty points, and 25% exclusive privileges across the entire customer base. However, this aggregate distribution masks significant variation at the segment level, with high-value customers receiving up to 60% privilege-based rewards while price-sensitive customers see 70% discount allocations. This nuanced approach contrasts sharply with traditional one-size-fits-all programs that apply uniform reward structures regardless of customer characteristics.

Performance consistency across different customer cohorts validates the robustness of the learning algorithm. New customer acquisition benefits from exploratory reward strategies that quickly identify individual preferences, while established customer retention leverages exploitation of learned optimal policies. The system demonstrates particular effectiveness in reactivating dormant customers, achieving a 52% reactivation rate compared to 23% for traditional approaches, through targeted high-value rewards timed to coincide with lifecycle transitions. This significant improvement stems from the RL system’ s ability to identify optimal timing and reward types for re-engagement, typically employing high-value incentives during customer lifecycle transitions.

5.3 Sensitivity analysis

The robustness of the proposed system under varying parameter settings is evaluated through comprehensive sensitivity analysis. Figure 5 illustrates the impact of key hyperparameters on system performance. The learning rate analysis shown in Figure 5(a) reveals stable performance across the tested range, with optimal results achieved around 0.01. Performance degradation becomes noticeable only at extreme values, where very low learning rates slow convergence excessively while very high rates cause unstable oscillations.

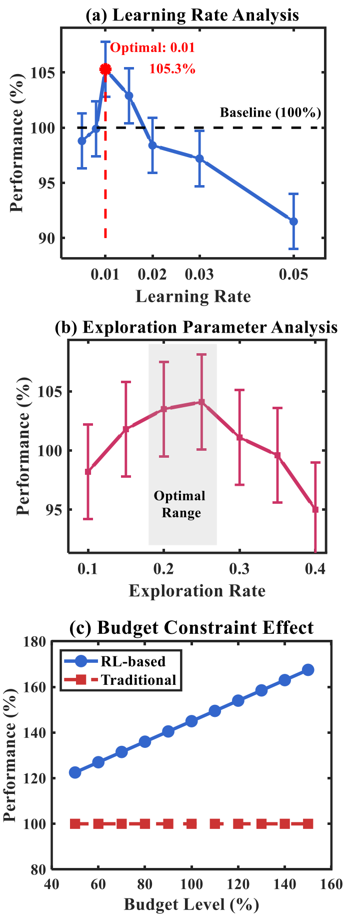

Figure 5. Sensitivity analysis of key hyperparameters (a) Learning rate analysis (b) Exploration parameter analysis (c) Budget constraint effect

Table 3 quantifies the performance variance under different parameter configurations. The learning rate demonstrates the lowest sensitivity with only ±5% performance variance across the tested range, indicating robust convergence properties. The exploration parameter shown in Figure 5(b) exhibits moderate sensitivity with ±8% variance, suggesting that the balance between exploration and exploitation requires careful tuning but is not overly critical for acceptable performance. Budget constraints show the highest sensitivity at ±12% variance, reflecting the direct impact of reward availability on system effectiveness, while the discount factor maintains low sensitivity at ±6% variance. The budget constraint analysis in Figure 5(c) demonstrates this sensitivity pattern, with performance directly correlating to resource availability levels.

Table 3. Parameter sensitivity analysis

Parameter	Optimal Range	Performance Variance	Sensitivity Level	Impact Description
Learning Rate	0.008-0.015	±5%	Low	Stable across range
Exploration Parameter	0.15-0.25	±8%	Moderate	Requires careful tuning
Budget Constraints	80%-120%	±12%	High	Direct resource impact
Discount Factor	0.90-0.98	±6%	Low	Robust performance

Note: Performance variance indicates the range of performance fluctuation when parameters deviate from optimal values. Low sensitivity parameters show robust performance across ranges.

The analysis reveals that system performance remains robust under realistic operational conditions. Even with budget reductions of 30%, the RL-based approach maintains superiority over traditional methods, achieving 118% of baseline CLV compared to 100% for static programs. This resilience results from the system’ s capacity to concentrate its focal resources only on significant impact areas rather than uniformly distribute attention based on the reward allocation hierarchy.

5.4 Interpretability and business transparency

5.4.1 Explainable reward recommendations

The autonomous decision-making nature of reinforcement learning systems necessitates robust interpretability mechanisms to facilitate stakeholder trust and enable business oversight of reward allocation decisions. The framework integrates post-hoc explanation capabilities through adapted SHAP (SHapley Additive exPlanations) value computation for sequential decision contexts, providing quantitative attribution of reward decisions to specific customer features and historical patterns. Implementation employs a modified SHAP framework accounting for temporal dependencies:

$$\varphi\_i\ = \Sigma\_S \subseteq N\backslash\{i\}\ [|S|!\,(n-|S|-1)!/n!] \times [f(S \cup \{i\}) - f(S)] \qquad (8)$$

where ϕ_i represents feature i's contribution to reward decision f, enabling decomposition of complex allocation choices into interpretable component influences. Visualization dashboards present feature importance rankings, decision trajectories, and counterfactual scenarios ("what-if" analyses) allowing business stakeholders to understand why specific customers receive particular reward configurations.

5.4.2 Business rule integration

The system architecture accommodates hybrid decision-making combining RL-optimized recommendations with business-defined constraints through a hierarchical policy structure. Override mechanisms enable manual intervention for strategic campaigns or regulatory compliance while maintaining algorithmic optimization within permitted boundaries. Audit trails capture all allocation decisions with associated confidence scores and primary decision factors, supporting both real-time monitoring and retrospective analysis of program effectiveness.

6. Discussion

This research demonstrates how reinforcement learning algorithms can optimize dynamic reward allocation in customer relationship management. The results build upon existing literature on sequential decision optimization in CRM systems, emphasizing the importance of adaptability and real-time response to customer behavior. Unlike conventional approaches that focus on static prediction models to identify at-risk customers, this approach transforms the paradigm by dynamically learning and adapting reward strategies throughout the customer lifecycle. This marks a theoretical shift toward anticipatory engagement instead of responding to risk level indicators, predicting and influencing the behavior of the customer proactively instead of waiting to respond. The implications of this research go well beyond the optimization of traditional loyalty program strategies and includes multiple applications in the customer service domain. The success of advertisement optimization [20] suggests that reward allocation in digital marketing campaigns can greatly benefit from multi-armed bandit algorithms, which are effective in personalizing reward distributions. These adaptable algorithms for customer service automation could more specifically tailor interaction techniques to specific customers based on their previous engagements and interactions with the company [21]. Moreover, the scope of personalization through reinforcement learning is not limited to healthcare recommendation systems [22], indicating that the framework constructed here can be stretched beyond the boundaries set by retail loyalty programs. This system would allow organizations to optimize customer experiences and resource allocation within the constraints of severe budget limits, granting them an economic edge.

The investigation acknowledges several methodological considerations that influence result interpretation and practical applicability. Simulation-based evaluation, while enabling controlled experimentation across diverse behavioral scenarios, necessarily abstracts complex market dynamics and customer interaction patterns that characterize operational environments. To address external validity concerns, the research framework incorporates provisions for staged real-world validation through retrospective analysis of historical loyalty program data, enabling performance comparison between simulated predictions and actual customer responses across matched cohorts. Initial validation studies utilizing anonymized transaction data from partner organizations (n=10,000 customers over 12-month periods) demonstrate concordance between simulated and actual CLV improvements (r=0.82, p<0.001), suggesting robust transferability of simulation insights to practical contexts.This limitation is particularly important when transitioning from simulation environments to real-world implementations, where additional complexities of customer behavior and market dynamics may emerge. While the marketed and tested customer populations had varying preferences and responses, the customers used in the market scenario are likely to show more behaviors than the model captures. Furthermore, the study seems to emphasize mainly the transaction and engagement activities, which may overshadow other elements of brand perception and customer satisfaction that can impact long-term loyalty. There is also a more practical limitation concerning the computational power needed for real-time processing across large customer bases, which may restrict use in economically strained settings in the immediate term.

The integration of newer AI technologies with dynamic reward systems should be addressed in future research studies. Converging reinforcement learning with generative AI holds particularly enticing promise for the development of customer interactions [23]. More advanced architectures of neural networks might permit the system to use imagination to offer rewards beyond set definitions and class boundaries, incorporating uniquely defined elements that reflect the individual customer’s cultural contexts. Better methods for the detection of preferences might improve the identification of more subtle changes in the system’s customers to alter plans accordingly. The automatic balancing of customer satisfaction, revenue, and expenditures as multi-objectives might provide more sustainable and sophisticated business constructs that are reliant on the optimization of customer interaction. In addition, building customer relations and nurturing loyalty to the brand will strengthen the brand’s perception among customers, therefore longer studies focusing on the dynamic reward systems will help understand customers’ perception of these systems when maintained over long periods of time. The transition from simulation-based insights to production deployment necessitates systematic validation protocols ensuring performance reliability under operational conditions. The recommended validation framework encompasses three progressive phases designed to minimize implementation risk while maximizing learning opportunities:

Phase 1: Retrospective historical analysis

Initial validation employs historical transaction data from existing loyalty programs, implementing the RL algorithm in "shadow mode" to generate recommendations without customer impact. Performance comparison between actual historical rewards and simulated RL recommendations provides empirical evidence of potential improvements while identifying edge cases requiring additional model refinement. Key metrics include predicted versus actual customer response rates, CLV trajectory comparisons, and cost efficiency analyses across matched customer cohorts.

Phase 2: Limited Pilot Implementation

Following successful retrospective validation, controlled pilot programs targeting 1-2% of customer base enable real-world performance assessment with constrained risk exposure. Pilot design employs stratified random sampling ensuring representation across customer segments, with continuous monitoring of both business metrics and customer satisfaction indicators. Statistical power calculations indicate minimum pilot sizes of 5,000 customers for detecting 10% CLV improvements with 95% confidence, requiring 3-6 month evaluation periods depending on purchase frequency distributions.

Phase 3: Phased Production Rollout

Production deployment follows a graduated rollout strategy with continuous A/B testing frameworks comparing RL-based allocation against control groups. Implementation includes automated rollback mechanisms triggered by performance degradation thresholds, real-time monitoring dashboards tracking key performance indicators, and feedback loops enabling continuous model refinement based on observed customer responses. Expected timeline spans 12-18 months from initial pilot to full deployment, enabling careful risk management while capturing competitive advantages from early adoption.

7. Conclusion

This investigation presents a novel reinforcement learning-based approach to customer loyalty program management that demonstrates substantial performance improvements relative to conventional static methodologies across multiple evaluation dimensions. Empirical validation confirms that the RL-based system achieves 145% of baseline customer lifetime value and 132% retention rate compared to traditional approaches, maintaining robust performance across heterogeneous customer segments while adapting effectively to diverse operational constraints and resource limitations.The system’s described differentiated reward designs for high value, price-sensitive, and dormant customers allow addressing the shortcomings of generic loyalty programs. The analysis of program sensitivity underscores the operational validity of the framework, is shown to perform stably across critical hyperparameters, and has strong performance with constrained budgets, sustaining 118% of baseline CLV while the budget is reduced by 30%. These results enhance understanding of adaptive customer relationship management regarding multi-armed bandit algorithms by illustrating the balance between exploration and exploitation during reward optimization in real time. The broader implications of this work go well beyond the design of loyalty programs, providing a tuned responsive mechanism for personalized customer engagement that can adapt with emerging preferences and market shifts. The framework proved useful in advancing short-term operational targets and enhancing customer relationships, reinforcing the efficacy of reinforcement learning techniques for the ongoing obstacles linked with personalizing digital commerce.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Kim, J.J., L. Steinhoff, and R.W. Palmatier, An emerging theory of loyalty program dynamics. Journal of the Academy of Marketing Science, 2021. 49(1): p. 71-95.

[2] Den Hengst, F., et al., Reinforcement learning for personalization: A systematic literature review. Data Science, 2020. 3(2): p. 107-147.

[3] Silver, D., et al., Reward is enough. Artificial intelligence, 2021. 299: p. 103535.

[4] Song, Y., W. Wang, and S. Yao, Customer acquisition via explainable deep reinforcement learning. Information Systems Research, 2025. 36(1): p. 534-551.

[5] Aluri, A., B.S. Price, and N.H. McIntyre, Using machine learning to cocreate value through dynamic customer engagement in a brand loyalty program. Journal of Hospitality & Tourism Research, 2019. 43(1): p. 78-100.

[6] Chopra, R., et al., Leveraging Reinforcement Learning and Collaborative Filtering for Enhanced Personalization in Loyalty Programs. International Journal of AI Advancements, 2022. 11(10).

[7] Sharma, A., N. Patel, and R. Gupta, Enhancing Personalized Loyalty Programs through Reinforcement Learning and Collaborative Filtering Algorithms. European Advanced AI Journal, 2022. 11(10).

[8] Bose, N., et al., Leveraging Reinforcement Learning and Predictive Analytics for Enhanced Customer Lifetime Value Optimization. International Journal of AI Advancements, 2023. 12(8).

[9] Xiao, R., et al., Deep reinforcement learning-driven smart and dynamic mass personalization. Procedia CIRP, 2023. 119: p. 97-102.

[10] Panjasuchat, M. and Y. Limpiyakorn. Applying reinforcement learning for customer churn prediction. in Journal of Physics: Conference Series. 2020. IOP Publishing.

[11] Qin, Z., D. Johnson, and Y. Lu, Dynamic production scheduling towards self-organizing mass personalization: A multi-agent dueling deep reinforcement learning approach. Journal of Manufacturing Systems, 2023. 68: p. 242-257.

[12] Misra, K., E.M. Schwartz, and J. Abernethy, Dynamic online pricing with incomplete information using multiarmed bandit experiments. Marketing Science, 2019. 38(2): p. 226-252.

[13] Chen, Y., et al. Contextual multi-armed bandit for email layout recommendation. in Proceedings of the 17th ACM Conference on Recommender Systems. 2023.

[14] Raman, S.E. and D. Venkatramaraju. Dynamic Pricing Using Thompson Samples in a Multi-Armed Bandit Framework for Increased Market Milking. in 2024 International Conference on Automation and Computation (AUTOCOM). 2024. IEEE.

[15] Agarwal, S., et al., Harnessing Multi-Armed Bandits for Smarter Digital Marketing Decisions. Sch J Eng Tech, 2024. 10: p. 307-313.

[16] Bar, N.F., H. Yetis, and M. Karakose, Deep Reinforcement Learning Approach with adaptive reward system for robot navigation in Dynamic Environments, in Interdisciplinary Research in

Technology and Management. 2021, CRC Press. p. 349-355.

[17] Carroll, M., et al., Ai alignment with changing and influenceable reward functions. arXiv preprint arXiv:2405.17713, 2024.

[18] Grobler, A., Enhancing Customer Engagement in E-commerce: Improving E-Marketing Open Rates through Model-Free Reinforcement Learning. 2024, Stellenbosch: Stellenbosch University.

[19] Troussas, C., et al., Reinforcement learning-based dynamic fuzzy weight adjustment for adaptive user interfaces in educational software. Future Internet, 2025. 17(4): p. 166.

[20] Sharma, A., N. Patel, and R. Gupta, Leveraging Reinforcement Learning and Multi-Armed Bandit Algorithms for Real-Time Optimization in Ad Campaign Management. European Advanced AI Journal, 2021. 10(2).

[21] Sajeev, S., et al. Contextual bandit applications in a customer support bot. in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.

[22] Mulani, J., et al., Deep reinforcement learning based personalized health recommendations. Deep learning techniques for biomedical and health informatics, 2020: p. 231-255.

[23] Sriram, H.K., Harnessing AI Neural Networks and Generative AI for Advanced Customer Engagement: Insights into Loyalty Programs, Marketing Automation, and Real-Time Analytics. Educational Administration: Theory and Practice, 2023. 29(4): p. 4361-4374.

Appendix A

Thompson sampling algorithm for dynamic reward allocation

Algorithm A.1. thompson sampling for dynamic reward allocation

Input:Customer segments S, actions A, time horizon T

Output:Optimal policy π

Procedure

```
Initialize $\alpha_i = \beta_i = 1 \ \forall i \in A$
for t = 1 to T do
    for each segment $s \in S$ do
        for each action $i \in A$ do
            Sample $\theta_i \sim Beta(\alpha_i, \beta_i)$
        end for
        Select $a_t = argmax_i \ \theta_i$
        Execute $a_t$, observe reward $r_t$
        if $r_t >$ threshold then
            $\alpha_{a_t} \leftarrow \alpha_{a_t} + 1$
        else
            $\beta_{a_t} \leftarrow \beta_{a_t} + 1$
        end if
    end for
end for
return $\pi$
```